70638

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 15 /42

A2

(11) International Publication Number: **WO 97/40462**

(43) International Publication Date: 30 October 1997 (30.10.97)

(21) International Application Number: PCT/US97/06457

(22) International Filing Date: 18 April 1997 (18.04.97)

(30) Priority Data:
08/636,517 19 April 1996 (19.04.96) US

(71) Applicant: SPECTRA BIOMEDICAL, INC. [US/US]; 4040 Campbell Avenue, Menlo Park, CA 94025 (US).

(72) Inventor: PEROUTKA, Stephen, J.; 1025 Tournament Drive, Hillsborough, CA 94010 (US).

(74) Agents: STORELLA, John, R. et al.; Townsend and Townsend and Crew L.L.P., 8th floor, Two Embarcadero Center, San Francisco, CA 94111 (US).

(81) Designated States: AU, CA, JP, KR, MX, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: CORRELATING POLYMORPHIC FORMS WITH MULTIPLE PHENOTYPES

(57) Abstract

This invention provides a database containing value sets indicating polymorphic forms and phenotypes for each member of a subject population, and methods of analyzing the database to determine the correlation between the polymorphic form at at least one genetic locus and at least two phenotypes.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

CORRELATING POLYMORPHIC FORMS WITH MULTIPLE PHENOTYPES

BACKGROUND OF THE INVENTION

This invention relates to the use of a digital computer to analyze a database. More specifically, the database contains medical information.

Over recent years, much progress has been made in mapping and sequencing the human genome, and it is to be expected that the full coding sequence of nearly every human gene will be known within a few years. However, determining the function of newly identified genes, or, conversely, identifying phenotypes associated with genes, has proceeded more slowly. Elucidation of function can be particularly difficult in situations in which a single gene contributes to several phenotypes and/or where multiple genes contribute to a single phenotype. Such situations may prove to be the norm rather than the exception. Some examples of genes correlating with multiple phenotypes include ApoE (heart disease and Alzheimer's disease), the AT gene (ataxis, telangiectasias, radiation sensitivity, leukemia, breast cancer and diabetes); the *brca1* gene (breast and ovarian cancer); Huntington's gene (movement disorder, dementia and psychosis); and the leptin receptor (diabetes, obesity). Some examples of diseases correlated with multiple genes include heart disease, Alzheimer's disease, hypertension, diabetes, and obesity.

Existing approaches to correlating genetic polymorphism and phenotype often start by selecting a single phenotype of interest (for example, diabetes, migraine or Alzheimer's disease). A population having the phenotype is selected together with a control population who lack the phenotype. DNA is extracted from both populations and co-segregational linkage between the phenotype and polymorphic markers in the DNA is performed. Usually, the analysis initially identifies polymorphic markers spaced some distance from the gene associated with the phenotype. By a variety of approaches, such as direct cloning, it is often possible to identify markers progressively closer to the gene until eventually the gene itself is identified.

Having found a variant form of a gene, or a polymorphic marker that correlates with a single disease phenotype, the above approach has, in some instances, been extended to look for correlations with one or more additional phenotypes. The approach in identifying a correlation with a second phenotype has been very similar to that for the first phenotype. That is, a further population of individuals is identified that have the second phenotype. This population typically has entirely different individuals from the population having the first phenotype. One then tests for a correlation between the variant gene or polymorphic marker and the population having the second phenotype in comparison with a control population.

Existing approaches toward correlating a genetic defect with multiple phenotypes are illustrated by work performed to identify phenotypic correlations with a polymorphism in human complement C3 ("C3"). The complement system contains a number of serum proteins that play a major role in immunologically-mediated inflammation (McLean and Winklestein (1984) *J. Pediatr.* 105:179-188). C3 is an acute phase reactant active in both the classic and alternative complement pathway. Its synthesis is increased during acute inflammation. Its fragments (C3a and C3b) have anaphylatoxic, chemotactic and histaminic actions and affect smooth muscle function. Activation of the complement system is usually beneficial to the host, as in resistance to infections, but may also lead to cytolysis of "normal" cells, as in autoimmune disorders. A common polymorphism in the C3 gene (C3F) that results from a single base pair change (M. Botto et al. (1990) *J. Exp. Med.*, 172:1011-1017) exists in Caucasoid populations (gene frequency = 0.20) (Weime and Demeulenaere (1967) *Nature* 214:1042-1043; Alper and Propp (1968) *J. Clin. Invest.* 47:2181-2191).

Sorensen and co-workers have investigated an association between the C3F allele and three disease phenotypes; multiple sclerosis, atherosclerosis and hypertension (Jans & Sorenson (1980) *Acta Neurol. Scandinav.* 63:237-342; Sorensen & Dissing (1975) *Human Heredity* 25:279-283; Schaadt et al. (1981) *Clin. Sci.* 61:363-365). In each study, populations having and lacking the respective disease symptoms were identified, and the individuals in the populations were tested for the presence of C3F and C3S alleles. In each study, it was reported that the C3F allele occurred significantly more frequently in individuals from the populations having the disease symptoms than in the control populations lacking disease symptoms. Although each of these studies was

performed by the same research group, the populations used to study the correlation between C3F and each phenotype appear to have been entirely different.

Several additional studies investigating possible correlations of C3F with the above and other disease phenotypes have been performed by a variety of research groups. The frequency of the C3F polymorphism has been reported to be increased in common disorders such as essential hypertension (Kristensen and Petersen (1978) *Circulation* 58:622-625; Schaadt et al. (1981) *Clin. Sci.* 61:363s-365s) and chronic polyarthritis (Farhud et al. (1972) *Humangenetik* 17:57-60; Bronnestam (1973) *Hum. Hered.* 23:206-213; Puttick et al. (1990) *Ann. Rheum. Dis.* 49:225-228. In addition, the frequency of C3F has also been reported to be increased in a variety of less common disorders such as bronchial asthma (Srivastava et al. (1985) *Hum. Hered.* 35:263-264), chronic renal failure (Reguerio and Arnaiz-Villena (1984) *Hum. Genet.* 67:437-440), renal allograft dysfunction (Andrews et al. (1995) *Transplantation* 60:1342-1346), Crohn's Disease (Elmgreen et al. (1984) *Acta. Med. Scand.* 215:375-378), hepatitis (Farhud et al. (1972) *Humangenetik* 17:57-60), autoantibody nephritic factor (Finn and Matieson (1993) *Clin. Exp. Immunol.* 91:410-414), IgA nephropathy (Rambausek et al. (1987) *Nephrol. Dial. Transplant* 2:208-211), Indian childhood cirrhosis (Srivastava & Srivastava (1984) *Hum. Hered.* 35:268-270), type 2 membranoproliferative glomerulonephritis (McLean & Winklestein (1984) *J. Pediatr.* 105:179-188), juvenile onset systemic lupus erythematosus (McLean & Winklestein (1984) *J. Pediatr.* 105:179-188) and systemic vasculitis (Finn et al. (1994) *Nephrol. Dial. Transplant* 9:1564-1567). No consistent association of C3F has been found with ulcerative colitis, leprosy, diabetes, or hyperlipidemia (Farhud et al. (1972) *Humangenetik* 17:57-60; Puttick et al. (1990) *Ann. Rheum. Dis.* 49:225-228.

Despite these various reported correlations between C3F and disease phenotypes, a causative role for C3F in disease remains controversial. For example, Welch et al., (1990) *J. Pediatr.* 226:92-7 reported that a biochemical comparison of purified forms of the C3F and C3S proteins did not identify any significant differences in functional properties between the two. The authors noted that many of the previous correlation studies had yielded results of marginal significance and some had subsequently been questioned. The authors concluded that further epidemiological surveys of C3 genetics in specific diseases were not warranted.

SUMMARY OF THE INVENTION

In one aspect this invention provides a method performed in a programmable digital computer. The method comprises providing a database having, for each member of a subject population; a) a first value set specifying at least one polymorphic form at at least one genetic locus exhibiting polymorphism, and b) a second value set specifying a plurality of phenotypes, wherein at least one of said polymorphic forms is not known to have a statistically significant correlation with at least one of said phenotypes; and determining the statistical correlation between the at least one polymorphic form and the plurality of phenotypes. In one embodiment of the invention, the step of providing the database comprises providing a nucleic acid sample for each member and determining the at least one polymorphic form from the nucleic acid samples. The database is, preferably, a relational database.

In another aspect, the invention provides the above-described database.

In another embodiment the invention provides a kit comprising a database having, for each member of a subject population, a value set specifying a plurality of phenotypes, and a DNA sample from each member of the subject population.

In another aspect, the invention provides a method for determining the degree of risk that a subject has or will develop a phenotype or syndrome. The method involves determining whether the subject has a polymorphic form shown by the method of this invention to have a statistically significant correlation with the phenotype or syndrome. Having the polymorphic form indicates the positive or negative risk of developing the phenotype or syndrome.

In another aspect, the invention provides a method of determining whether a human subject is at increased risk of allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease comprising determining whether the subject has a C3F allele, whereby the presence of the allele indicates that the subject is at increased risk.

In another aspect, the invention provides a kit comprising at least one nucleic acid probe capable of detecting in a subject a polymorphic form identified by the method of this invention as having a significant positive statistical correlation with a phenotype or syndrome, and instructions indicating that the presence of the polymorphic form indicates that the subject is at positive risk of developing the phenotype or syndrome.

In another aspect, this invention provides a kit comprising at least one nucleic acid probe for distinguishing between C3F and C3S alleles; and instructions indicating that the C3F allele confers susceptibility to allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease. In one embodiment, the kit comprises two probes capable of acting as primers for amplification of genomic DNA around nucleotide 364 of exon 3 of the C3F gene allele.

In another aspect, this invention provides a method of treating a subject having allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease associated with the C3F syndrome, comprising administering an immunosuppressive agent or an anti-histamine to the subject.

In another aspect, this invention provides a kit comprising an immunosuppressive agent or an anti-histamine and instructions for use of the immunosuppressive agent or anti-histamine the treatment of allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease associated with C3F syndrome.

In another aspect, this invention provides a method for screening a compound for the ability to inhibit C3F protein. The method involves contacting a C3F protein with the compound; and determining whether the compound inhibits C3F-mediated complement activity. A compound that inhibits the activity of C3F is a candidate drug in the treatment of C3F syndrome.

In another aspect, this invention provides a computer program product for analyzing a database, in particular a relational database, comprising: code that receives as input a database having, for each member of a subject population; a) a first value set specifying at least one polymorphic form at least one genetic locus exhibiting polymorphism, and b) a second value set specifying a plurality of phenotypes, wherein at least one of said polymorphic forms is not known to have a statistically significant correlation with at least one of said phenotypes; code that determines the statistical correlation between the at least one polymorphic form and the plurality of phenotypes; and a computer readable medium that stores the codes. In one embodiment, the program product further comprises code that displays results of the statistical analyses. In another embodiment, the computer program product further comprises code that displays results of the statistical analyses. When the database contains information on a variety of polymorphic forms, a programmer can select the one or ones to analyze.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates an example of a computer system used to execute the software of the present invention. Fig. 1 shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons such as mouse buttons 13. Cabinet 7 houses a CD-ROM drive 15 and a hard drive (not shown) that may be utilized to store and retrieve computer programs including code incorporating the present invention. Although a CD-ROM 17 is shown as the computer readable storage medium, other computer readable storage media including floppy disks, DRAM, hard drives, flash memory, tape, and the like may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

Fig. 2 shows a system block diagram of computer system 1 used to execute the software of the present invention. As in Fig. 1, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 102, system memory 104, I/O controller 106, display adapter 108, removable disk 112, fixed disk 116, network interface 118, and speaker 120. Removable disk 112 is representative of removable computer readable media like floppies, tape, CD-ROM, removable hard drive, flash memory, and the like. Fixed disk 116 is representative of an internal hard drive, DRAM, or the like. Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 102 (i.e., a multi-processor system) or memory cache.

Arrows such as 122 represent the system bus architecture of computer system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, display adapter 108 may be connected to central processor 102 through a local bus or the system may include a memory cache. Computer system 1 shown in Fig. 2 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art. In a preferred embodiment, the computer system is a work station from, e.g., Sun Microsystems.

Fig. 3 presents a flow chart for performing the analyses of this invention.

DETAILED DESCRIPTION OF THE INVENTION

I. DEFINITIONS

The term "polymorphic form" refers to one of the forms of a polymorphism existing at a genetic locus in a population of organisms. At the molecular level, when two homologous segments of genetic material have different nucleotide sequences, the segments exhibit polymorphism and each different sequence is a polymorphic form. An example of polymorphic forms is the different sequences for a gene that encode polypeptides of different amino acids sequences, or functional RNA molecules having different sequences. The differences may reflect nucleotide substitutions, insertions or deletions between the different sequences. Another example of polymorphic forms is a variable number of tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats and tetranucleotide repeats. Polymorphic forms can be identified by different genetic markers such as restriction patterns for DNA, e.g., restriction fragment length polymorphisms ("RFLPs"). Polymorphic forms also manifest as different mendelian alleles for a gene. A polymorphic form can be identified within a sub-group of a population sharing a common characteristic, e.g., sex (male or female), classification (e.g., primates).

The term "phenotype" refers to any distinguishable trait of an organism. A phenotype can be a physical trait or a mental (e.g., emotional) trait. Physical traits include both normal (non-disease) physical variants and disease states. Disease states include, for example, general medical conditions (e.g., heart disease, cancer, autoimmune diseases, inflammatory disorders, diabetes), neurological conditions (e.g., Alzheimer's disease, migraine) and psychiatric conditions (e.g., anxiety, depression). Non-disease phenotypes include, for example, hair color, left- or right-handedness, height and baldness. Phenotypes used in a database typically include several conditions selected from these different groupings.

Determination of the existence of a phenotype in an individual depends on the phenotype. Some phenotypes are obvious from inspection (e.g., baldness, eye color). Many phenotypes can be elicited from questionnaires (e.g., diseases such as migraine, anxiety, depression). Some phenotypes can also be determined by referring to prior medical or other records (e.g., cancer, diabetes, heart attack). Some phenotypes are best determined by biochemical analysis of tissue samples (e.g., enzymatic deficiencies). Phenotypes are usually defined by conventional criteria (e.g., Diagnostic and Statistical

Manual), but can be defined by customized criteria depending on the intended use of the database. The value that specifies the phenotype may be binary, i.e., an individual does or does not have the phenotype. Alternatively, the value may be scaler, such as height, weight, or blood pressure, or a selected specific trait, such as eye color.

In a preferred embodiment, the data on the phenotypes are collected by conducting a thorough medical work-up and history of the subjects in the population. This can involve identifying all past and present illnesses and complaints, both mental and physical, as well as physiological tests on a variety of systems, including, for example, blood work-ups, cardiovascular tests, etc.

The term "statistical correlation" refers to a statistical association between two values as measured by any statistical test including, for example, chi-squared analysis, ANOVA or multivariate analysis. The correlation between a polymorphic form and a phenotype is considered statistically significant at a P value of < 0.05, and, preferably < 0.01 or < 0.001.

The term "syndrome" refers to a collection of phenotypes having a statistically significant correlation with a polymorphic form. The polymorphic form is said to be "associated" with the phenotypes.

The term "subject" refers, preferably, to humans, but also to mammals and other animals, other multicellular organisms such as plants, single celled organisms or viruses.

The term "nucleic acid probe" refers to a nucleic acid molecule which binds to a specific sequence or sub-sequence of another nucleic acid molecule. A probe is preferably a nucleic acid molecule which binds through complementary base pairing to the full sequence or to a sub-sequence of a target nucleic acid. It will be understood by one of skill in the art that probes may bind target sequences lacking complete complementarity with the probe sequence depending upon the stringency of the hybridization conditions. The probes are preferably directly labelled as with isotopes, chromophores, lumiphores, chromogens, or indirectly labelled such as with biotin to which a streptavidin complex may later bind. By assaying for the presence or absence of the probe, one can detect the presence or absence of the select sequence or sub-sequence.

A "label" is a composition detectable by spectroscopic, photochemical, biochemical, immunochemical, or chemical means. For example, useful labels include ^{32}P, fluorescent dyes, electron-dense reagents, enzymes (e.g., as commonly used in an

ELISA), biotin, dioxigenin, or haptens and proteins for which antisera or monoclonal antibodies are available (e.g., by incorporating a radio-label into the peptide, and used to detect antibodies specifically reactive with the peptide). A label often generates a measurable signal, such as radioactivity, fluorescent light or enzyme activity, which can be used to quantitate the amount of bound label.

A "labeled nucleic acid probe" is a nucleic acid probe that is bound, either covalently, through a linker, or through ionic, van der Waals or hydrogen bonds to a label such that the presence of the probe may be detected by detecting the presence of the label bound to the probe.

The term "code" refers to human-readable source code, which is the instructions written by the programmer in a programming language, as well to executable machine code, which is the instructions of a program that are converted from source code to instructions that the computer can understand.

The term "relational database" means a database that stores information in tables -- rows and columns of data -- and conducts searches by using data in specified columns of one table to find additional data in another table.

II. MULTIPHENOTYPIC ANALYSIS

The invention provides novel methods of using a programmable digital computer for determining the statistical correlation between a polymorphic form of interest and each of multiple phenotypes. Correlations are achieved through the provision of a database, which stores a matrix of information. The database is created from a subject population. Each individual in the population is tested for the presence of a plurality of phenotypes. The database indicates whether each individual has or lacks each of the phenotypes. Preferably, the database is a relational database.

A. Selecting The Population

The number of individuals in the population is selected so as to yield a statistically meaningful measure of the correlation between the genotypic polymorphism and the phenotypes to be tested. This number, in turn, depends on the frequency of the polymorphic form in the population, the frequency of the phenotype in the population, and the power of the statistical test being used for make the correlation. For example, if the chi-squared test is used for two polymorphic forms, each at a frequency of about

50%, and two phenotypes, each at a frequency of about 50%, then each analysis cell needs about five data points, or twenty data points in all, for meaningful statistical analysis. If a polymorphic form is present at a frequency greater than about 10%, most statistical analyses will require a population of at least about 200 members. If the polymorphic form is found in about 1% to 10% of the population, the test population needs to have at least about 1000 members. If the polymorphic form is found in about 0.1% to 1% of the population, the test population in the database requires at least about 10,000 members. Thus, the population can have at least 50, at least 100, at least 200, at least 500, at least 1000 at least 5000 or at least 10,000 members.

The population of individuals to be the subject of the database is often selected without prior knowledge or testing of the phenotypes of any individual. For example, the population can be selected at random. The population can be selected at random with respect to the phenotypic traits. Alternatively, the population can be selected to reflect a diversity of phenotypes. Also, the population can be directed with respect to a particular phenotype, for example, 50% with the trait and 50% without. In some databases, populations comprise unrelated individuals. For example, no two individuals can be from the same nuclear family (i.e., two parents and children of those parents.) In other databases, some individuals may be from the same family.

B. Selecting The Polymorphic Forms

The database also indicates, for each individual, the polymorphic form at at least one genetic locus exhibiting polymorphism and the identity of multiple phenotypes. The polymorphic form typically is determined from DNA testing, although polymorphic forms determined from mendelian analysis also are useful. In the case of DNA testing, for each individual included in the database, a sample of DNA, usually genomic DNA, is obtained. This sample is often obtained at about the same time that phenotypes are being determined for that individual. Typically, the DNA is placed in storage until a later time when a candidate genotypic marker for investigation becomes available. Thus, at the time at which phenotypes are being determined one typically does not know the genotypic marker(s), which will be correlated with the phenotypes in the database. In one embodiment, the invention provides a database including value sets specifying a plurality of phenotypes for each subject of the population and a DNA sample for each member of the population indexed to identify the subject from whom it came.

Often, one locus is analyzed at which two polymorphic forms have been identified. However, if there are many polymorphic forms at a locus, several can be tested. Optionally, polymorphic forms from several loci can be tested. Sometimes the polymorphic form occurs within a gene or putative gene of unknown function, for example, a newly sequenced open reading frame. For example, the polymorphic form may represent a variation preventing expression of a functional gene product. Other times the polymorphic form occurs with a gene that has been characterized for a biochemical function (e.g., binding or enzymatic activity) but for which a correlation with phenotype has not been determined. Other times the genotypic marker occurs within a gene that has been correlated with certain phenotypes, but for which it is believed that correlations with additional phenotypes can be found. Other times, the polymorphic form has no known statistically significant correlation with a disease phenotype. In other situations, the genotypic marker is not itself within a gene or non-coding sequence of any significance, but is linked to such a sequence. The human genome project is detecting many genes whose function is unknown. In one embodiment, the polymorphic form is for a gene of unknown function.

Having selected a locus of interest, the polymorphic form for each subject in the population at the locus is determined. Naturally, heterozygotes will have two polymorphisms. The polymorphic forms are added to the existing database.

C. Computer Analysis

The computer is programmed to perform an analysis of the database to determine the statistical correlation between a polymorphic form at at least one locus and at least two phenotypes. The chosen polymorphic form typically does not have a known statistical correlation with at least one of the phenotypes tested. Statistical correlation is typically determined by the chi-squared method. The computer can specifically identify correlations that are statistically significant. Both negative and positive statistically significant correlations are of interest. Different phenotypes can be ranked by the strength of correlation.

A typical database usually contains information on at least about 10, 20, 50 or 100 phenotypes, most of them disease phenotypes. In one embodiment, the number of subjects in the population is at least 100 and the number of phenotypes analyzed is at least 20. Some of the phenotypes included in the database are known or

suspected to exhibit co-morbidity with each other (e.g., hypertension and heart disease), whereas other phenotypes lack known co-morbidity with each other. In one embodiment, the phenotypes selected each involve a plurality of tissue types or organ systems, e.g., nervous system, circulatory system, digestive system, connective tissue, etc. In one embodiment, the population is selected independently of the phenotypes, i.e., without regard for the presence or absence of any particular phenotypes.

Multiple genotypic markers can be correlated with phenotypes by a simple extension of the above approach. For example, if mutations A and B occur at different places within the same gene, one might analyze whether the presence of A or B correlates with disease. If mutations A and B occur in different genes, which are suspected to have a similar function, one might analyze whether the presence of A and B shows stronger correlation with disease phenotype than A or B alone. Large numbers of polymorphic forms (e.g., 5, 10, 20 or 100) can be analyzed in different combinations in this manner.

In one aspect, this invention provides a computer program product for analyzing a database that includes code that receives as input the database code that receives as input at least one selected polymorphic form; code that performs statistical analyses on the selected polymorphic form and phenotypes in the database; and a computer readable medium that stores the codes. In certain embodiments, the computer program product further includes code that displays results of the statistical analyses. In another embodiment, the computer program product further includes code that displays results of the statistical analyses. It may, for example, select and display all statistically significant correlations. The database may be stored in the hard drive or a floppy disk and moved into the computer memory as needed to perform the analysis. The product can include code that receives as input instructions from a programmer, e.g., for identifying polymorphic forms or phenotypes to be correlated, or the form or contents of the display.

In one embodiment, the relational database has separate tables with value sets specifying general medical history, presence or absence of migraine, psychiatric history, subject identifiers, and personal information. The tables are linked by a record number which also acts as an index.

III. DIAGNOSTIC METHODS

A. General

The result of this analysis is a list in which the degree of correlation between one or more polymorphic forms and a plurality of phenotypes is determined. This information is valuable in a number of ways.

A statistically significant positive or negative correlation between a polymorphic form and a disease phenotype indicates that with respect to the general population, a subject has a positive or negative risk of developing the disease. In general, the greater the statistical correlation, the greater the positive or negative risk of developing the disease. Accordingly, this invention provides methods of determining whether a person is at positive or negative risk of developing a disorder, trait, or syndrome. The method involves selecting a polymorphic form identified by the method of this invention to have a statistically significant correlation with a phenotype and determining whether the subject has the polymorphic form.

When the method uncovers a statistically significant correlation between one or more phenotypic forms and several phenotypes, these phenotypes constitute a syndrome of traits that are associated with each other. Identification of syndromes is useful designing modes of prophylaxis and treatment, and in uncovering the common source of the syndrome.

The subject or his or her treating physician can be informed if the subject has a polymorphic form positively correlated with a phenotype or with a syndrome. Patients possessing such a polymorphic form can often be counselled as to the risk of contracting the syndrome and about relatively benign prophylactic measures to take before the onset of disease. For example, the risk of developing heart disease, and cancer can be reduced by changes in diet and life style. Other diseases can be prevented by taking low doses of drugs. For example, migraine, anxiety and depression can be prevented by low dose prophylactic treatment with SSRIs. A further benefit of the analysis is that it identifies genes that should be given priority for further analysis. That is, genes containing genotypic markers showing strong correlations with common human diseases. The functions of these genes can be analyzed in vitro or in transgenic animal models, and compounds can be identified that antagonize or agonize the function of gene products.

B. C3F Syndrome

As described in the Examples, the methods of this invention demonstrate that the C3F polymorphic form has a significant statistical correlation with allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease. Thus, these diseases, as well as others already known to be related to C3F, represent a syndrome associated with the presence of the C3F allele. The term "C3F syndrome" as used herein refers to disease state in which a person exhibits several or all of the disorders associated with C3F and has a C3F allele. Persons at positive risk of developing C3F syndrome can be identified by determining whether the person has the C3F allele. The person can be informed that they have the allele and that they are at increased risk of contracting one or more of the associated diseases. The persons also can be counseled on prophylactic measures to take to decrease the risk of contracting an associated disease. For example, the subject can be counseled to change diet. Also, an immunosuppressive agent can be administered to the subject.

The C3F allele can be detected genetically by, for example, isolating genomic DNA from the subject and detecting a variation at nucleotide 364 in exon 3 of a C3 gene. A method for doing this is described in M. Botto et al. (1990), *supra*. Briefly, the C3F and C3S variants differ in the presence of an HhaI site at nucleotide 364 of exon 3. Primers based on the nucleotide sequence of the gene are used to amplify a stretch of genomic DNA including this site. Exposure to HhaI will or will not cleave the amplified fragment, depending upon whether or not the restriction site is present. The cleaved and uncleaved fragments are detectable by, for example, agarose gel electrophoresis.

IV. DIAGNOSTIC KITS

The kits of this invention include nucleic acid probes for detecting in a subject a polymorphic form identified by the methods of this invention as having a significant positive statistical correlation with a phenotype or syndrome, in particular with a disease phenotype, and instructions indicating that the presence of the polymorphic form indicates that the subject is at positive risk of developing the phenotype or syndrome. Normally, the probe will be labeled and will be useful for hybridizing with a restriction fragment on, for example, a Southern blot, thereby identifying the polymorphic form.

In one embodiment, the kit contains at least one nucleic acid probe for detecting the C3F allele. Such probes can, for example, amplify genomic DNA around nucleotide 364 of exon 3. Examples of such primers are Ex3 (SEQ ID NO:1) and Ex2 (SEQ ID NO:2).

V. THERAPEUTIC KITS

C3F syndrome is treatable prophylactically or therapeutically by administering immunosuppressive drugs and/or anti-histamines. Accordingly, this invention provides a kit comprising an immunosuppressive agent and/or an antihistamine and instructions to use the drug in the treatment of allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease, disorders of C3F syndrome associated for the first time with C3F by determining their statistical correlation using a database as described herein.

VI. METHODS OF SCREENING

This invention also provides methods of screening for an agent effective to treat C3F syndrome. The methods comprise contacting a C3F protein with a compound and determining whether the compound inhibits C3F-mediated complement activity. Inhibition indicates that the compound is a candidate as a drug to treat the syndrome. In one embodiment, the C3F protein is contacted with the compound in a mixture containing cells, an antibody that specifically binds to the cells, and complement proteins C1, C2, and C4-C9. Complement activity is determined from lysis of the cells.

The following examples are offered by way of illustration, not by way of limitation.

EXAMPLE

I. METHODS

A. Clinical Genetic Relational Database

Individuals were recruited for participation in a long term clinical genetic study of common medical disorders. All individuals were interviewed by a physician, nurse and/or trained clinical investigator, and all clinical interviews were reviewed by a physician. Subjects were evaluated using a semi-structured interview which included

questions about demographics, personal medical history, family medical history, a detailed review of systems, medication use, and healthcare utilization. Informed consent was obtained and DNA samples collected. A random sample of 200 DNA samples from unrelated Caucasians, who were 35 years of age or older, was selected for analysis from a pre-established clinical genetic database. All clinical data were obtained independently of the genotypic data.

B. Genotyping

Genotyping of C3 using the *HhaI* RFLP polymorphism was performed in the study participants. Genomic DNA was isolated using the Puregene DNA isolation kit (Gentra Systems, Research Triangle Park, North Carolina). Genomic DNA was amplified and restriction enzyme digested as described by M. Botto et al. (1990) *J. Exp. Med.* 172:1011-1017. Briefly, 1 μg of genomic DNA was amplified using 25 pM of primer Ex3 (5' ATCCCAGCCA ACAGGGAG 3' (SEQ ID NO:1)) and Ex4 (5' TAGCAGCTTG TGGTTGAC 3' (SEQ ID NO:2)), corresponding to positions 328-345 and complementary to nucleotides 514-531[30], respectively.

The amplification was performed in 25 μL of a buffer containing 10 mM Tris, pH 8.3, 1.5 mM $MgCl_2$, 50 mM KCl, 200 μM each dNTP and 1 U *Taq* polymerase (Perkin Elmer, Foster City, CA). The reaction was initially denatured at 95°C for 4 minutes, followed by 30 cycles of amplification with denaturation at 95°C for 1 minute, annealing at 56°C for 1 minute, extension at 72°C for 1 minute and a final extension at 72°C for 10 minutes. 10 U of HhaI (New England Biolabs, Beverly, MA) was added directly to the amplification reaction and digested at 37°C for greater than two hours. The products were separated on a 1.5% agarose SFR gel (Amresco, Solon, OH). Analysis of the polymorphic form was performed by two individuals blinded to the clinical status.

II. RESULTS

A. General Population Survey

Genotyping of the 200 person sample set indicates that 4.5% of individuals display the C3F/C3F polymorphic form, 31.5% have the C3S/C3F polymorphic form and 64% have the C3S/C3S genotype. These data are consistent with the Hardy-Weinberg prediction. In the current dataset, the C3F gene frequency is 0.20 and the

C3S gene frequency is 0.80. These values are consistent with the C3 allele frequencies reported in Caucasoid populations (Cavalli-Sforza, et al. The history and geography of human genes, (Princeton University Press, Princeton, NJ, 1994). The average age of the study participants is 54 $\pm$ 1 years in the C3F positive group and 53 $\pm$ 1 years in the C3F negative group. No significant differences in gene frequencies were found between males and females.

B. Clinical Genetic Relational Database Screening

The frequency of the C3F allele was then analyzed in a variety of common clinical disorders by screening a clinical genetic database. As shown in Table 1, a number of disorders are significantly more frequent in C3F positive individuals. Food allergies, arthritis, coronary heart disease (CHD), hypertension with CHD, osteoporosis, and thyroid disease are significantly more frequent in the C3F positive individuals than in the C3F negative individuals. Hypertension is increased in the C3F positive group but the increase does not reach statistical significance.

C. Food Allergies

Food allergies (defined as symptoms of immediate type hypersensitivity reactions to shellfish, vegetables or fruit) were reported by 42 of the 200 individuals in the present study. Urticaria and angioedema were reported by 11 individuals following shellfish ingestion and by 31 individuals following fruit or vegetable ingestion. The frequency of food allergies is 36% (i.e. 25 of the 70 individuals for whom a diagnosis was made) in the C3F positive group and 13% (n = 17) in the C3F negative group, a highly significant difference in frequency (Chi-square = 13.41; $p < 0.0001$). The C3F gene frequency in this group of food allergy sufferers is 0.33, a frequency which is significantly greater (Chi-square = 11.81; $p < 0.0003$) than the C3F frequency (0.18) in subjects without food allergies.

D. Non-rheumatoid Arthritis

Symptoms of chronic non-rheumatoid arthritis were reported by 53 of the 200 individuals in the present study. All 53 individuals had been diagnosed independently with non-rheumatoid arthritis by their personal physician. The frequency of arthritis is 38% (n = 27) in the C3F positive group and 21% (n = 26) in the C3F

negative group, a significant difference in frequency (Chi-square = 6.99; $p < 0.004$). The C3F gene frequency in this group of arthritis sufferers is 0.29, a frequency which is significantly greater (Chi-square = 7.16; $p < 0.004$) than the C3F frequency (0.17) in subjects without arthritis.

E. Coronary Heart Disease (CHD) and Hypertension

CHD, defined as angiographically documented coronary artery disease, was reported by 16 of the 200 individuals in the present study. The frequency of CHD is 13% (n = 9) in the C3F positive group and 6% (n = 7) in the C3F negative group. The difference in CHD frequency between the C3F positive and C3F negative group reaches statistical significance (Chi-square = 3.03; $p < 0.04$). A similar pattern is observed with hypertension. Hypertension is reported by 60 of the individuals in the present study to have been diagnosed by their personal physician. The frequency of hypertension is 35% (n = 25) in C3F positive group and 28% (n = 35) in the C3F negative group. The difference in hypertension frequency between the C3F positive and C3F negative groups does not reach statistical significance. When the analysis for CHD is restricted to only hypertensive patients in the study sample (i.e., 60 individuals), the incidence of CHD is 32% (n = 8) of the C3F positive vs. 9% (n = 3) of the C3F negative individuals, a statistically significant difference in CHD frequency (Chi-square = 5.35; $p < 0.01$).

F. Osteoporosis

Osteoporosis was reported by 5 of the 200 individuals in the present study. All 5 individuals had been diagnosed independently by their personal physician and were being treated for osteoporosis at the time of the interview. 4 of the 5 individuals with osteoporosis are C3F positive. The C3F gene frequency in the osteoporosis group is 0.60, a frequency which is significantly greater (Chi-square = 9.95; $p < 0.0008$) than the C3F frequency (0.19) in subjects without osteoporosis.

G. Thyroid Disease

Hypothyroidism was reported by 27 of the 200 individuals in the present study. All 27 individuals had been diagnosed independently by their personal physician and 25 had been treated with thyroid replacement. The frequency of hypothyroidism is

19% (n = 14) in the C3F positive group and 10% (n = 13) in the C3F negative group. The difference in hypothyroidism frequency between the C3F positive and C3F negative group reaches statistical significance (Chi-square = 3.32; $p < 0.03$). The C3F gene frequency in the thyroid disease group is 0.28, a frequency which is greater than the C3F frequency (0.19) in subjects without thyroid disease.

III. DISCUSSION

These results indicate that the presence of a C3F allele is associated with multiple clinical phenotypes which can be attributed to tissue injury that occurs secondary to a functional hyperactivity of the complement system. In contrast to previous studies which focused on specific disease groups, this invention utilizes a novel approach to genetic analysis by using a database to characterize the clinical relevance of the C3F allele. Significant associations were observed between C3F and food allergies, non-rheumatoid arthritis, CHD, hypertension with CHD, osteoporosis and thyroid disease. Hypertension was also increased in C3F positive individuals. Thus, the correlations discovered in this invention together with previous data support the hypothesis that activation of the complement system in C3F positive individuals, in comparison to C3F negative individuals, can lead to an increase in endothelial damage and secondary arteriosclerosis.

The method of this invention identified a number of novel clinical associations. For example, a highly significant increase in food allergies were reported by the C3F positive vs. C3F negative individuals. Food allergy is an immunologically mediated response to an ingested food antigen. Symptoms may consist of nausea, vomiting, diarrhea, urticaria, angioedema, bronchospasm and/or rhinitis. Reactions to food antigens are more pronounced and/or more common in C3F positive individuals compared to C3F negative individuals.

Evidence for tissue damage secondary to immune system hyperactivity in C3F positive individuals is also consistent with the present observation that an increased frequency of thyroid disease and osteoporosis is present in C3F positive individuals. Chronic inflammatory thyroid disease (or Hashimoto's Disease) is a common disorder of middle age in which autoimmune factors are believed to play a prominent role. Osteoporosis is a common disorder of elderly women in which mast cells and

macrophages are believed to play a role. These results strongly support the conclusion that tissue injury is increased in C3F positive compared to C3F negative individuals.

In addition, C3F has been reported to be increased in a wide variety of less common disorders (Table 2). The diseases in which the C3F frequency has been found to be increased are similar in that the immune system and/or inflammation is believed to play a prominent role in the pathophysiology of each disorder. The underlying molecular mechanism that might clinically link these various disorders has yet to be elucidated. Based on the data presented here, it is suggested that tissue injury secondary to a functional hyperactivity of the complement system in C3F positive individuals could be the molecular mechanism that underlies all of these otherwise unrelated disorders. Theoretically, the increased efficacy of C3F following an immunological challenge may lead to individually subtle, but cumulatively significant, increases in vascular lesions over the course of many years. C3F, in comparison to C3S, may lead to either increased local membrane damage or increased permeability of the endothelium following activation of the complement system. This conclusion is consistent with the present data which indicate that the clinical morbidity resulting from the C3F allele occurs over the course of many years.

As a result, the present data indicate that either direct or indirect inhibition of C3 offers a novel therapeutic approach to the prevention of chronic arthritis, hypertension, CHD and other disorders in as many as 20% of the general population. In addition, more aggressive life-long treatment of common immune-mediated inflammatory disorders such as hay fever, drug reactions and contact dermatitis decrease the long term risk from mild inflammatory arthritis, hypertension and CHD. Early and effective therapeutic interventions are a novel approach to decrease the incidence of a variety of chronic clinical disorders in C3F positive individuals.

TABLE 1

REPORTED FREQUENCIES OF VARIOUS CLINICAL DISORDERS IN A SAMPLE OF 200 HEALTHY ADULTS

Disorder	Percent Affected		p value
	C3F Positive (n = 72)	C3F Negative (n = 128)	
Food allergies	36%	13%	0.0001
Arthritis (non-rheumatoid)	38%	21%	0.004
Hypertension with CHD	32%	9%	0.01
Osteoporosis, arthritis	6%	1%	0.02
Thyroid disease	19%	10%	0.03
Coronary heart disease	13%	6%	0.04
Hypertension	35%	28%	n.s.

TABLE 2

LITERATURE REVIEW OF C3F FREQUENCIES IN VARIOUS DISORDERS

Disorder	C3F Frequency	Ratio of C3F+ to C3F- in Affecteds vs. Unaffecteds	p value
Atherosclerotic vascular disease	.264	1.6	0.0005
Atherosclerotic vascular disease	.272	1.6	0.0005
Bronchial asthma	.059	3.3	0.001
Indian childhood cirrhosis	.149	12	0.001
Essential hypertension	.264	1.9	0.003
Chronic polyarthritis (rheumatoid factor +)	.252	1.3	0.01
Hepatitis	.307	1.6	0.02
Untreated hypertensives	.206	1.8	0.03
Chronic renal failure	.425	1.9	0.03
Systemic vasculitis	.290	2.6	0.03
Rheumatoid arthritis	.250	1.5	0.03
Crohn's disease	.226	1.3	0.05
Multiple sclerosis	.275	1.9	0.05
Nephritic factor	.327	2.1	0.05
Treated hypertensives	.159	1.4	n.s.
Rheumatoid arthritis	.232	1.2	n.s.
Mild inflammatory arthritis	.294	1.3	n.s.

The present invention provides a novel method for determining the correlation between polymorphic forms and phenotypes. While specific examples have been provided, the above description is illustrative and not restrictive. Many variations of the invention will become apparent to those of skill in the art upon review of this specification. The scope of the invention should, therefore, be determined not with

reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.

All publications and patent documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent document were so individually denoted.

WHAT IS CLAIMED IS:

1. A method performed in a programmable digital computer comprising:

providing a database having, for each member of a subject population;

a) a first value set specifying at least one polymorphic form at at least one genetic locus exhibiting polymorphism and

b) a second value set specifying a plurality of phenotypes, wherein at least one of said polymorphic forms is not known to have a statistically significant correlation with at least one of said phenotypes; and

determining the statistical correlation between the at least one polymorphic form and the plurality of phenotypes.

2. The method of claim 1 wherein the database is a relational database.

3. The method of claim 1 wherein the population is a human population.

4. The method of claim 1 wherein the population comprises at least 100 subjects.

5. The method of claim 1 wherein the population comprises at least 200 subjects.

6. The method of claim 1 wherein the population comprises at least 1000 subjects.

7. The method of claim 1 wherein the population comprises at least 10,000 subjects.

8. The method of claim 1 wherein the plurality of phenotypes is at least 10 phenotypes.

9. The method of claim 1 wherein the plurality of phenotypes is at least 100 phenotypes.

10. The method of claim 1 wherein the phenotypes include heart disease, a cancer, a disease of the immune system, and a neuropsychiatric disease.

11. The method of claim 1 wherein the plurality of phenotypes includes a neurological condition, a psychiatric condition, or a general medical condition.

12. The method of claim 1 wherein at least one of the phenotypes is a disease phenotype.

13. The method of claim 1, wherein the disease phenotypes comprise at least first and second disease phenotypes lacking co-morbidity with each other.

14. The method of claim 1, wherein the population has at least 100 subjects and the plurality of phenotypes is at least 20.

15. The method of claim 1, wherein the population is selected independently of the phenotypes.

16. The method of claim 1 wherein the polymorphic form is identified by a restriction fragment length polymorphism pattern.

17. The method of claim 1 wherein the polymorphic form has a known statistically significant correlation with a disease phenotype.

18. The method of claim 1 wherein the polymorphic form has no known statistically significant correlation with a disease phenotype.

19. The method of claim 1 wherein the at least one polymorphic form is a plurality of polymorphic forms.

20. The method of claim 19 wherein the plurality is at least 10.

21. The method of claim 1 wherein the phenotypes are collected independently of the polymorphic forms.

22. The method of claim 1 wherein the step of providing the database comprises providing a nucleic acid sample for each member and determining the at least one polymorphic form from the nucleic acid samples.

23. The method of claim 1, wherein at least one polymorphic form is for a gene of unknown function.

24. The method of claim 1 further comprising identifying correlations that are statistically significant.

25. A database having, for each member of a subject population:

a) a first value set specifying at least one polymorphic form at at least one genetic locus exhibiting polymorphism, and

b) a second value set specifying a plurality of phenotypes, wherein at least one of said polymorphic forms is not known to have a statistically significant correlation with at least one of said phenotypes.

26. A kit comprising a database having, for each member of a subject population, a value set specifying a plurality of phenotypes, and a DNA sample from each member of the subject population.

27. A method for determining the degree of risk that a subject has or will develop a phenotype or syndrome comprising determining whether the subject has a polymorphic form shown by the method of claim 1 to have a statistically significant

correlation with the phenotype or syndrome; whereby having the polymorphic form indicates the positive or negative risk of developing the phenotype or syndrome.

28. The method of claim 27 wherein the phenotype or syndrome has a statistically significant positive correlation with the polymorphic form and whereby having the polymorphic form indicates that the person is at positive risk of developing the disease phenotype.

29. The method of claim 28 wherein the subject is a human.

30. The method of claim 29 further comprising counseling the subject about the risk of developing the phenotype or syndrome.

31. A method of determining whether a human subject is at increased risk of allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease comprising identifying whether the subject has a C3F allele, whereby the presence of the allele indicates that the subject is at increased risk.

32. The method of claim 31, further comprising informing the subject or a treating physician that the subject has, or is susceptible to allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease.

33. The method of claim 31, wherein determining comprises detecting in DNA from the subject a variation at nucleotide 364 in exon 3 of a C3 gene indicating the presence of the C3F allele.

34. The method of claim 31, wherein the subject is determined to have a C3F allele and the method further comprises counseling the patient to change diet.

35. The method of claim 31, wherein the subject is determined to have a C3F allele and the method further comprises administering an immunosuppressive agent and/or an anti-histamine to the subject.

36. A kit comprising at least one nucleic acid probe capable of detecting in a subject a polymorphic form identified by the method of claim 1 as having a significant positive statistical correlation with a phenotype, and instructions indicating that the presence of the polymorphic form indicates that the subject is at positive risk of developing the phenotype.

37. The kit of claim 36 wherein the phenotype is a disease phenotype.

38. The kit of claim 26 wherein the probe is an amplification primer.

39. A kit comprising at least one nucleic acid probe for distinguishing between C3F and C3S alleles; and instructions indicating that the C3F allele confers susceptibility to allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease.

40. The kit of claim 39 comprising two probes capable of acting as primers for amplification of DNA around nucleotide 364 of exon 3 of the C3F gene.

41. A method of treating a subject having allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease associated with the C3F syndrome, comprising administering an immunosuppressive agent or an anti-histamine to the subject.

42. A kit comprising an immunosuppressive agent or an anti-histamine and instructions for use of the immunosuppressive agent or anti-histamine the treatment of allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease associated with a C3F allele exhibiting allergy to foods, thyroid disease, osteoporosis, arthritis and/or heart disease.

43. A method of screening a compound for the ability to inhibit C3F protein comprising:

contacting a C3F protein with the compound; and

determining whether the compound inhibits C3F-mediated complement activity;

whereby a compound that inhibits the activity of C3F is a candidate drug in the treatment of C3F syndrome.

44. The method of claim 43, wherein the C3F protein is contacted with the compound in a mixture containing cells, an antibody that specifically binds to the cells, and complement proteins C1, C2, and C4-C9, and complement activity is determined from lysis of the cells.

45. A computer program product for analyzing a database comprising:

code that receives as input a database having, for each member of a subject population;

a) a first value set specifying at least one polymorphic form at at least one genetic locus exhibiting polymorphism, and

b) a second value set specifying a plurality of phenotypes, wherein at least one of said polymorphic forms is not known to have a statistically significant correlation with at least one of said phenotypes; and the polymorphic form at at least one locus and a plurality of phenotypes for each member of a subject population;

code that determines the statistical correlation between the at least one polymorphic form and the plurality of phenotypes; and

a computer readable medium that stores the codes.

46. The computer program product of claim 45 further comprising code that displays results of the statistical analyses.

47. The computer program product of claim 45 further comprising code that displays results of the statistical analyses.

48. The computer program product of claim 45 wherein the database is a relational database.

49. The computer program product of claim 45 further comprising code that receives as input instructions from a programmer.

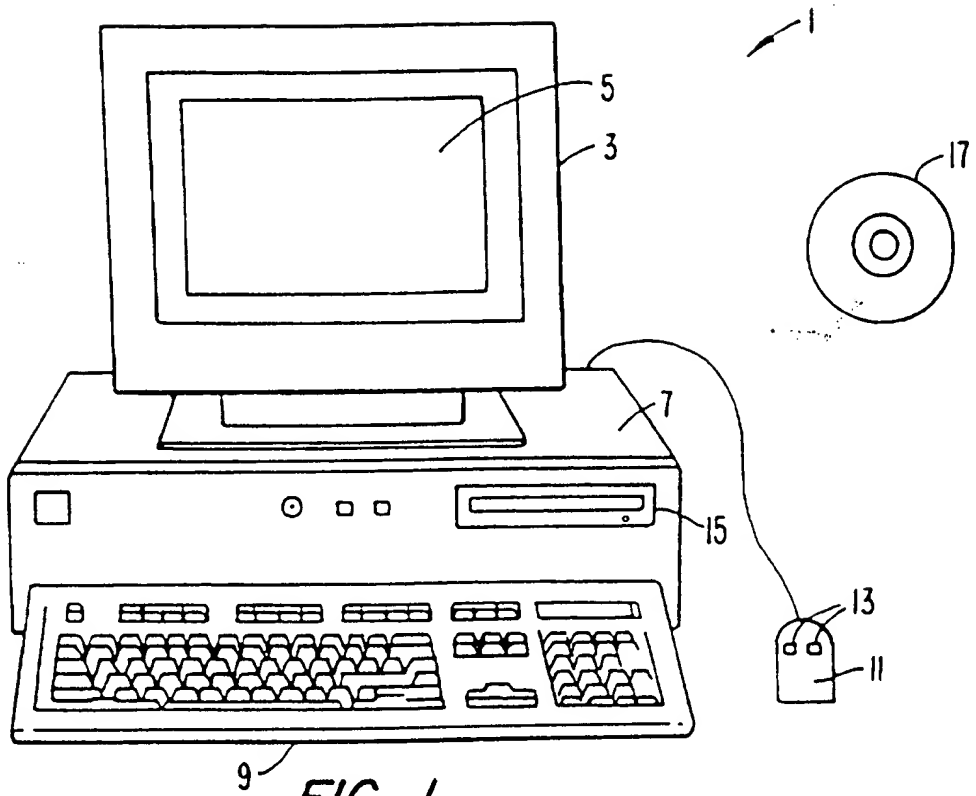
1
5
3
17
7
15
13
11
9

FIG. 1.

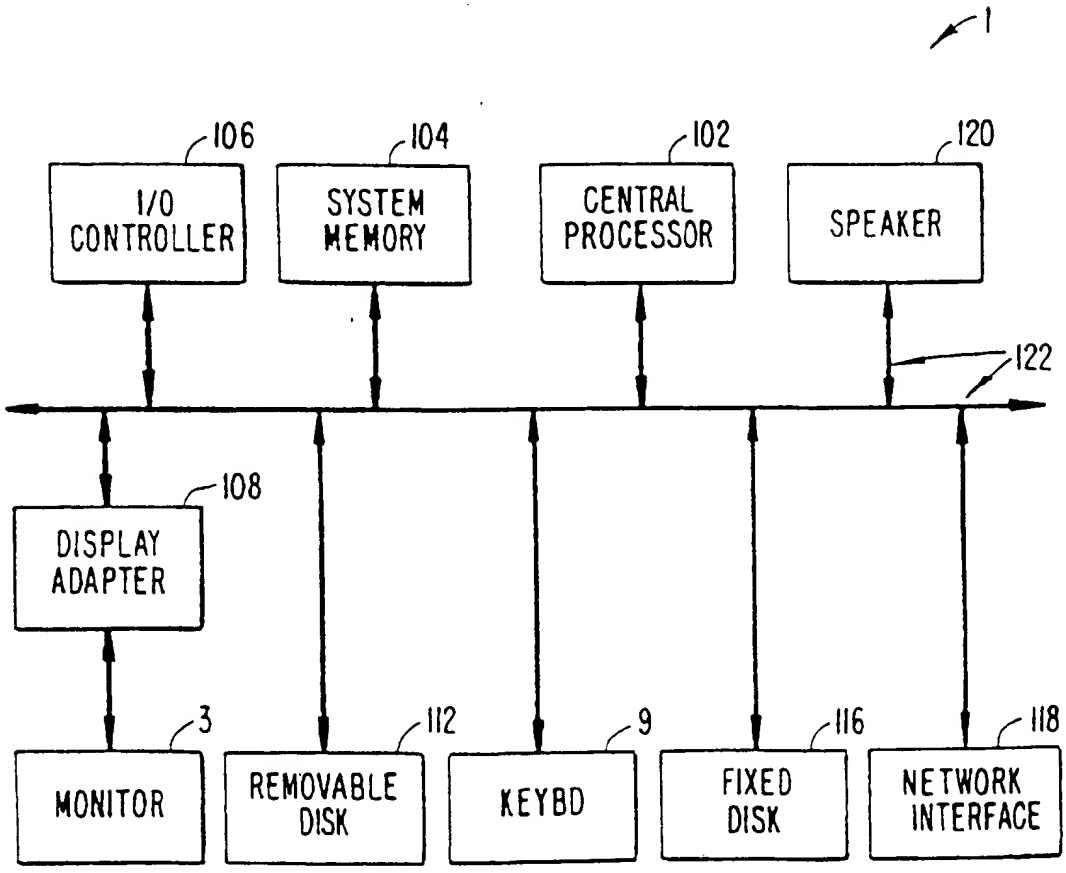
1
106
I/O CONTROLLER
104
SYSTEM MEMORY
102
CENTRAL PROCESSOR
120
SPEAKER
122
108
DISPLAY ADAPTER
3
MONITOR
112
REMOVABLE DISK
9
KEYBD
116
FIXED DISK
118
NETWORK INTERFACE

FIG. 2.

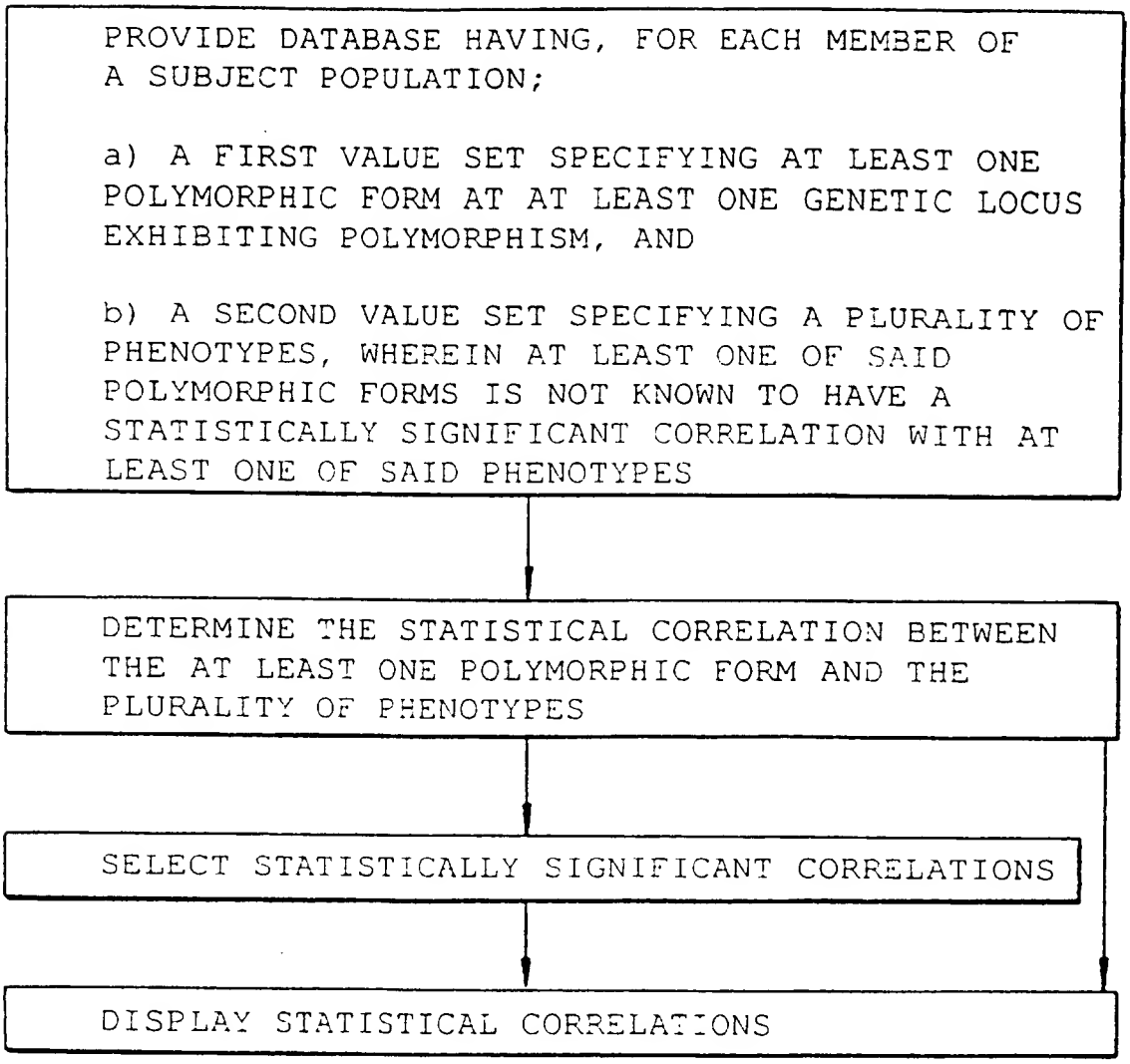
PROVIDE DATABASE HAVING, FOR EACH MEMBER OF A SUBJECT POPULATION;
a) A FIRST VALUE SET SPECIFYING AT LEAST ONE POLYMORPHIC FORM AT AT LEAST ONE GENETIC LOCUS EXHIBITING POLYMORPHISM, AND
b) A SECOND VALUE SET SPECIFYING A PLURALITY OF PHENOTYPES, WHEREIN AT LEAST ONE OF SAID POLYMORPHIC FORMS IS NOT KNOWN TO HAVE A STATISTICALLY SIGNIFICANT CORRELATION WITH AT LEAST ONE OF SAID PHENOTYPES
DETERMINE THE STATISTICAL CORRELATION BETWEEN THE AT LEAST ONE POLYMORPHIC FORM AND THE PLURALITY OF PHENOTYPES
SELECT STATISTICALLY SIGNIFICANT CORRELATIONS
DISPLAY STATISTICAL CORRELATIONS

FIG. 3.

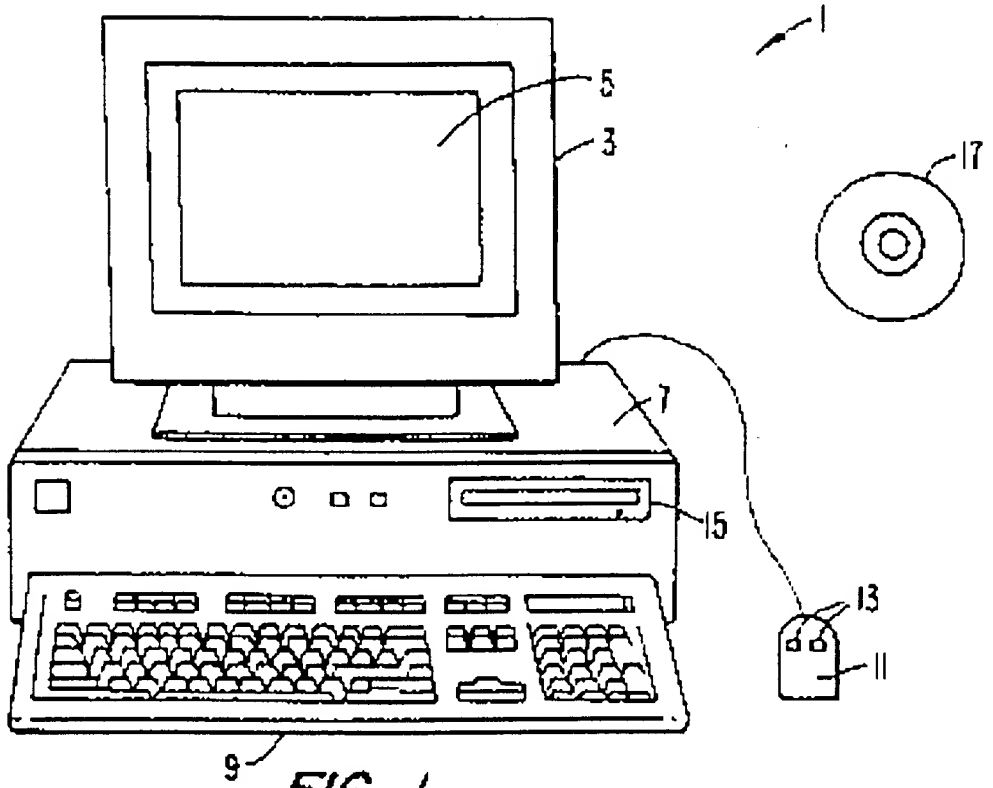
1
5
3
17
7
15
13
11
9

FIG. 1.

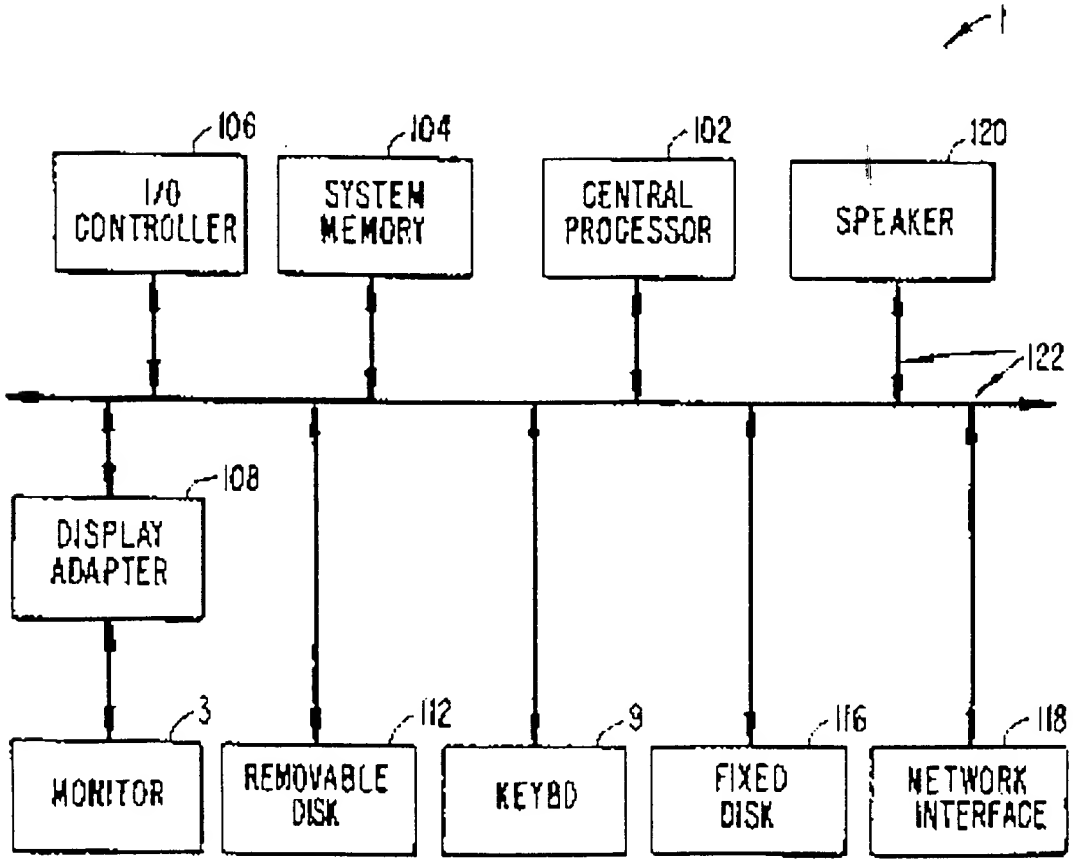
1
106
I/O CONTROLLER
104
SYSTEM MEMORY
102
CENTRAL PROCESSOR
120
SPEAKER
122
108
DISPLAY ADAPTER
3
MONITOR
112
REMOVABLE DISK
9
KEYBD
116
FIXED DISK
118
NETWORK INTERFACE

FIG. 2.

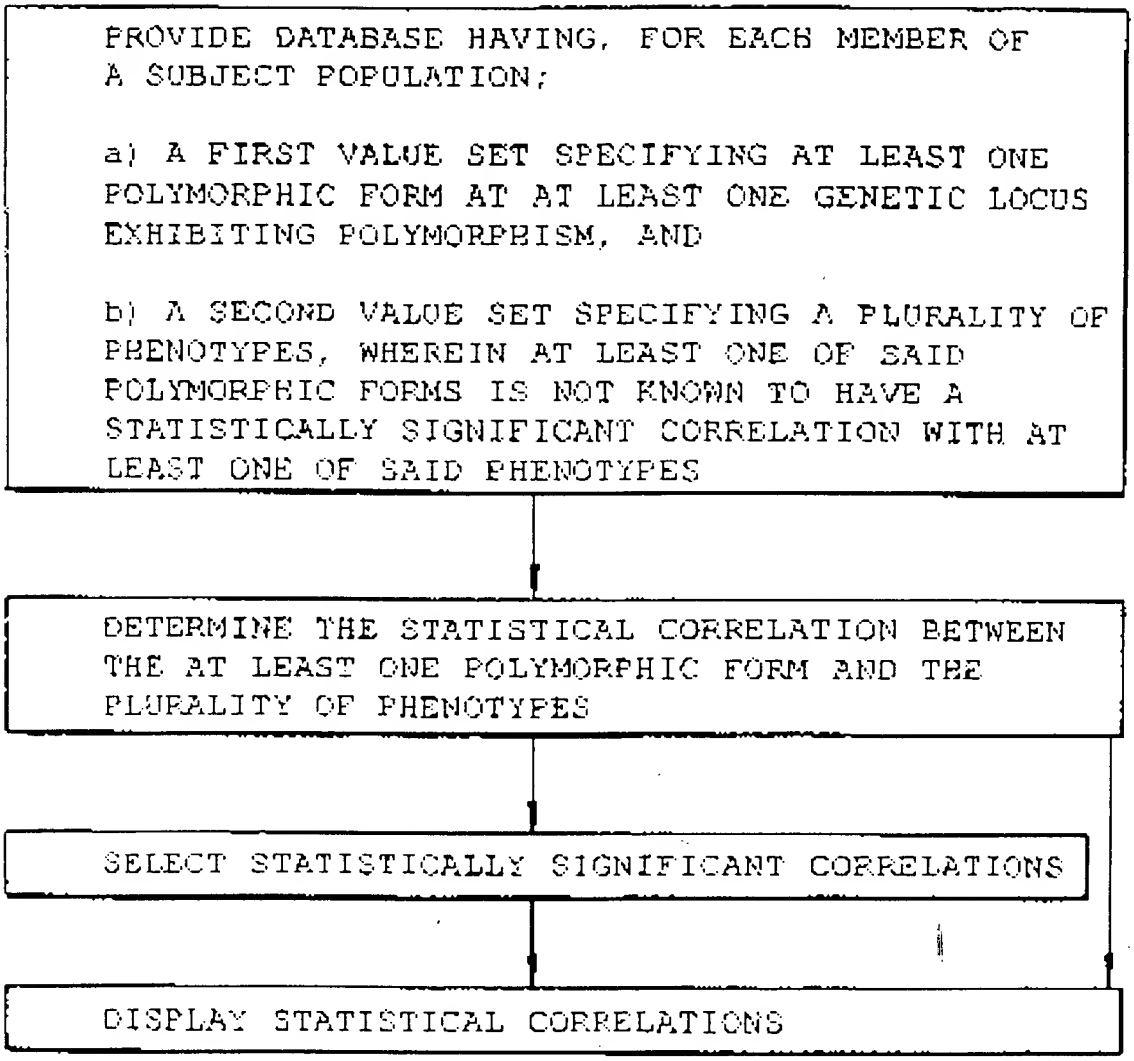
PROVIDE DATABASE HAVING, FOR EACH MEMBER OF A SUBJECT POPULATION;
a) A FIRST VALUE SET SPECIFYING AT LEAST ONE POLYMORPHIC FORM AT AT LEAST ONE GENETIC LOCUS EXHIBITING POLYMORPHISM, AND
b) A SECOND VALUE SET SPECIFYING A PLURALITY OF PHENOTYPES, WHEREIN AT LEAST ONE OF SAID POLYMORPHIC FORMS IS NOT KNOWN TO HAVE A STATISTICALLY SIGNIFICANT CORRELATION WITH AT LEAST ONE OF SAID PHENOTYPES
DETERMINE THE STATISTICAL CORRELATION BETWEEN THE AT LEAST ONE POLYMORPHIC FORM AND THE PLURALITY OF PHENOTYPES
SELECT STATISTICALLY SIGNIFICANT CORRELATIONS
DISPLAY STATISTICAL CORRELATIONS

FIG. 3.

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 19/00, C12Q 1/68, G01N 33/48, A61K 31/00, 38/00 // G06F 159:00

A3

(11) International Publication Number: **WO 97/40462**

(43) International Publication Date: 30 October 1997 (30.10.97)

(21) International Application Number: PCT/US97/06457

(22) International Filing Date: 18 April 1997 (18.04.97)

(30) Priority Data:
08/636,517 19 April 1996 (19.04.96) US

(71) Applicant: SPECTRA BIOMEDICAL, INC. [US/US]; 4040 Campbell Avenue, Menlo Park, CA 94025 (US).

(72) Inventor: PEROUTKA, Stephen, J.; 1025 Tournament Drive, Hillsborough, CA 94010 (US).

(74) Agents: STORELLA, John, R. et al.; Townsend and Townsend and Crew L.L.P., 8th floor, Two Embarcadero Center, San Francisco, CA 94111 (US).

(81) Designated States: AU, CA, JP, KR, MX, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(88) Date of publication of the international search report:
5 March 1998 (05.03.98)

(54) Title: CORRELATING POLYMORPHIC FORMS WITH MULTIPLE PHENOTYPES

(57) Abstract

This invention provides a database containing value sets indicating polymorphic forms and phenotypes for each member of a subject population, and methods of analyzing the database to determine the correlation between the polymorphic form at at least one genetic locus and at least two phenotypes.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 97/06457

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F19/00 C12Q1/68 G01N33/48 A61K31/00 A61K38/00
//G06F159:00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	COMPUTERS AND BIOMEDICAL RESEARCH, vol. 27, April 1994, US, pages 97-115, XP002039573 RITTER ET AL: "prototype implementation of the integrated genomic database" see page 97, line 5 - page 98, line 9 see page 99, line 39 - line 41 see page 103, line 10 - page 108, line 3 see page 111, line 16 - page 113, line 5; figures 1-6 ---	1-3,22, 23,25, 26,45, 48,49
A	INTERNATIONAL JOURNAL OF MAN-MACHINE STUDIES, vol. 23, no. 4, 1985, UK, pages 551-561, XP002039574 MUNAKATA: "knowledge-based systems for genetics" see page 552, line 29 - page 555, line 17 ---	1-3,25, 26,45, 48,49

-/--

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
3 September 1997	02.01.98
Name and mailing address of the ISA	Authorized officer

1

INTERNATIONAL SEARCH REPORT

Interna al Application No
PCT/US 97/06457

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	PROCEEDINGS OF THE 27TH SYMPOSIUM ON THE INTERFACE COMPUTING SCIENCE AND STATISTICS, 21 June 1995, US, pages 259-263, XP002039575 GALFALVY ET AL: "a statistical environment for multivariate analyses" ---	
P,X	COMPUTERS AND BIOMEDICAL RESEARCH, vol. 29, no. 4, August 1996, US, pages 327-337, XP002039576 CHEUNG ET AL: "phenodb : an integrated client/server database for linkage and population genetics" see the whole document -----	1-30, 36-39, 45-49

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 97/06457

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1 ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:

3 ☐ Claims Nos
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

SEE ANNEXED SHEET

1 ☐ As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims

2 ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee

3 ☐ As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos

4 ☒ No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-30, 36-38, 45-49

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest

☐ [illegible] the payment of additional search fees

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

1. Claims 1-30, 36-38, 45-49: A method performed in a programmable digital computer comprising a database, containing value sets indicating polymorphic forms and phenotypes for each member of a subject population. Methods for analyzing the database to determine the statistical correlations between a polymorphic form of interest and each of multiple phenotypes.

2. Claims 31-35,39-44: Methods and kits for determining increased risk for disease by identifying the presence of a C3F allele as well as methods for treating said diseases or for screening compounds inhibiting C3F protein.